View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Universiti Teknikal Malaysia Melaka: UTeM Open Journal System

Text Line Segmentation for Mushaf Al-Quran Using Hybrid Projection Based Neighbouring Properties

A. R. Radzid[1], M. S. Azmi[1], I. E. A. Jalil[1], N. A. Arbain[1], A. K. Draman @ Muda[1] and A. Tahir[2]
[1]*Faculty of Information and Communication Technology,*
Universiti Teknikal Malaysia Melaka, Malaysia.
[2]*Department of Information & Communication Technology,*
Politeknik Ungku Omar,
Jalan Raja Musa Mahadi,
31400 Ipoh, Perak, Malaysia.
sanusi@utem.edu.my

***Abstract*— Text line segmentation is an important step in document image processing. Its part of the pre-processing stage to prepared the images before throughout either feature extraction or classification images. In this paper, we present a method of line segmentation for Mushaf Al-Quran text using a hybrid projection based neighbouring properties. This is based on the pixel, object and histogram properties. This method will identify overlaps between neighbouring text lines and segment each line with precision. Overlap caused by interfering with diacritical marks or stroke of the Arabic word must be properly segmented without change the original meaning of the text. Experimental results show the validity of our method.**

***Index Terms*— Image Segmentation; Mushaf Al-Quran Documents; Text Line Segmentation.**

I. INTRODUCTION

Processing text document required text line segmentation and this is done during pre-processing stage. This is one of the important steps of Optical Character Recognition (OCR) and keyword spotting [1]. Line segmentation can be hard since the row of text can be overlapped on between neighbouring text lines cause by diacritical marks or stroke of the Arabic word. Overlapping is one of the problems during extract the text representation. It is also a problem of global text document processing and academic interest. Examples of overlap text can be found from document, manuscript, record, letters and many more.

Roman, Latin or English manuscript, document or handwritten differ from the Mushaf Al-Quran because Mushaf Al-Quran contains a lot of diacritical marks (Tashkil) that used as phonetic guides. This makes it harder to segment the line without accidentally cutting or misplacing diacritical marks or stroke of the Arabic word that will change the meaning of the word.

Overlapping may occur where space between neighbouring texts lines interfere with diacritical marks or stroke of the Arabic word when the parallel view is projected between neighbouring texts lines as shown in Figure 1. Thus, this will cause inaccuracy during line segmentation. This paper does not cover the issues of text when a character is completely connected on overlap spot.

This paper proposes a novel method in text line segmentation to identify overlaps between neighbouring text lines and segment each line with precision. The proposed method is using the technique of hybrid projection based neighbouring pixel, object and histogram properties.

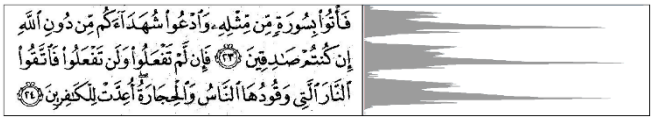


Figure 1: Result of horizontal projection histogram.

Text line segmentation has two categories of approaches. The first approach is a search for separating lines or paths. The second approach is a search for aligned physical units. It may depend on the complexity of text line structure of the document. Thus, projection profiles are commonly used for printed or offline document segmentation. This is because it can also be adapted to handwritten documents with little overlap.

II. RELATED WORK

The previous research relates to text line segmentation for Mushaf Al-Quran was done by Laith Nazeeh Bany Melhem [2] in 2015. His research successfully segments text line of Mushaf Al-Quran by using binary representation. In his research focusing on separate each text line and save in image form without any black space. He is using horizontal detection of each text line edges by calculating the white percentage for each line start from line 0 to Y-1. Base on his research, text line segmentation does not cover overlapping text line.

Projection profiles are commonly used for document segmentation. In 2004, Antonacopoulos and Karatzast [3] used projection profile for segmenting line of memorial/person record (World war II). While in 2003, He and Downton [4] used projection (RXY cuts) to segmenting line of Viadocs/Natural history cards. Moreover, in 2003, Pal and Datta [5] used piecewise projections for segmenting line for Indian handwritten documents. Next, in 2004, Zahour et al. [6] used piecewise projection and k-means clustering to segmenting line for ancient Arabic documents. Furthermore, in 1993, Shapiro et al [7] used projection profile to skewed separated lines for handwritten documents. Hence, this research focusing on domain of Mushaf Al-Quran text.

Xi Zhang and Chew Lim Tan [1] proposed line segmentation using constrained seam carving. They applied this technique in Greek, English and Indian. They proposed method was tried to extract all the text lines by computing the energy map only once. The problem with this techniques seems there no diacritics exits on their text test data. It will be a problem if applied their proposed method on Mushaf Al-Quran because Mushaf Al-Quran text contains a lot of diacritics.

Banumathi. K. L and Jagadeesh Chandra A P [8] proposed line segmentation using projection profile technique. They used Kannada handwritten text documents as data test. They applied horizontal projection profile, smoothening the projection, detect peaks and valleys and segmenting the line. This technique weak for test data Mushaf Al-Quran because Mushaf Al-Quran text contains overlapping cause by diacritics or Arabic word.

Fei Yin and Cheng-Lin Liu [9] proposed line segmentation using distance metric learning. They apply the proposed techniques on Chinese text line. Their novel text line segmentation algorithm based on minimal spanning tree (MST) clustering with distance metric learning. This technique weak for test data Mushaf Al-Quran because Mushaf Al-Quran text contains overlapping cause by diacritics or Arabic word.

This research is continuity from past paleography research topic. This research can relate with digital Jawi paleography field. This field of research can helps to identify authors, origin and date of manuscripts. By analyzing manuscript illumination can discover the information of specific manuscript [10]. Mohd Sanusi Azmi introduced features from triangle geometry for digit recognition of this field [11]. Moreover, this researcher applied his technique for Arabic or Jawi and this can be related for this research topic because of Al-Quran ware wrote in Arabic. Besides that, this researcher applied his techniques on Arabic calligraphy classification [12][13]. This calligraphy classification of the ancient manuscripts gives useful information to paleographers. Besides that, this research is also one of continuity pre-processing stage from the research of removing Al-Quran illumination [14]. This research is focusing on removing illumination from the text.

III. Proposed Method

A. Pre-processing

The main purpose of pre-processing is to enhance the inputted signal and to form images uniformly to be used throughout the process. Before performing this process, dataset must be prepared. Datasets that be used in this experiment are a collection of text images from Mushaf Al-Quran as an example are shown in Figure 2. Text image of Mushaf Al-Quran must not contain any decoration, illumination, illustration or any unnecessary object. Conventional steps such as noise removal and filtering include text normalization such as baseline correction, slant normalization and skew correction must be applied. These steps make the image process more reliable and effective [15].

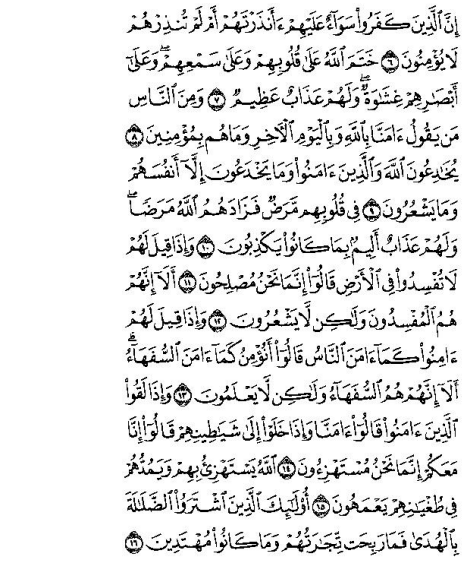


Figure 2: Example of text image Mushaf Al-Quran from Mushaf Al-Madinah Quran Majeed page 3.

Text image required in binarize form to easily perform to analysing each pixel of the text image. The text was first acquired from Mushaf Al-Quran and being transformed into a digitized image. Next, a uniform process to convert the image coloured pixel to grey-scale used the thresholding. Thresholding method that is used in this experiment was conducted by using Otsu's method proposed by Scholar Otsu in 1979 [16].

Thresholding method one of the important technique for image segmentation that converts a colour image into grey-scale. The basic idea of thresholding is to select an optimal grey-level threshold value for separating objects of interest in an image from the background based on their grey-level distribution. If g(x, y) is a threshold version of f(x, y) at some global threshold T, it can be defined as [6],

$$g(x, y) = 1 \text{ if } f(x, y) \geq T \quad (1)$$
$$= 0 \text{ otherwise}$$

Otsu's thresholding method was used in this research includes as one of the procedures for pre-processing. The Otsu method is a type of global thresholding in which it depends on only the grey value of the image. This method involves iterating through all the possible threshold values and calculating a measure of spread for the pixel levels on each side of the threshold [17]. Otsu's method defines the within-class variance as the weighted sum of the variances of each cluster [18].

After the image was converted to grey-scale then the image must be transformed into the binary colour based scale, the pixels were labelled as "1" for white pixels and black pixels as "0" (see Figure 3 and Figure 4).

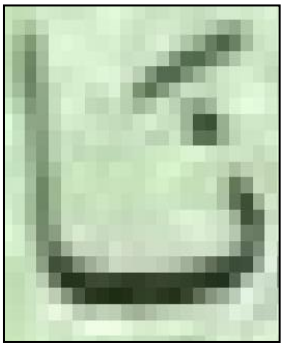


Figure 3: Example image of a word before Otsu's thresholding method.

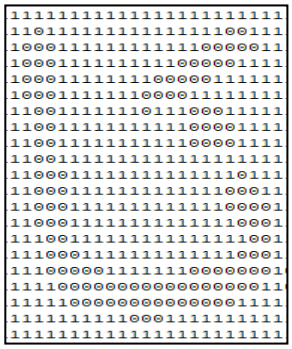


Figure 4: Example image of a word after Otsu's thresholding method and binarization.

Figure 3 showed an example of the input image and Figure 4 showed the result of the output image after pre-processing step was done.

B. Horizontal Projection Profile

This method will calculate each pixel by row to project its graph as shown in Figure 1. It only computes foreground pixel or text pixel as a value. By projecting the graph, we can determine peak contour frequency as a baseline while zero frequency as line space. Normally line space exits between two neighbouring texts lines which call upper text line and bottom text line. If zero frequency between peaks contours it specified as line space. But, if there is no zero frequency between peaks contours point at lowest contour specified as line space between peak contours. Then it determined as overlapping. It may contain diacritical marks or stroke of the Arabic word between neighbouring texts lines. Overview are illustrated in Figure 5. All pixel involve in this lower contour frequency will check its object possession based on its baseline.

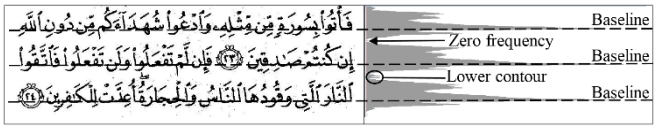


Figure 5: Overview of horizontal projection histogram.

C. Object Possession

This step is to make sure overlapping object either diacritical marks or stroke of the Arabic word are not misplaced by allocated in an incorrect row of the text line. The overlapping object is illustrated in Figure 6.

In order to fabricate the specified object, this research used neighbouring pixel properties. It will define and identify pixel based on the properties of its neighbour. It will continually cluster the pixel if the similarity criterion is statisfied. The neighbouring pixel that does not having similar properties will be refused from being clustered. This will continually iterate until there are no properties that are same as the specified pixels. The region that cover is surrounding the touching pixel point only.

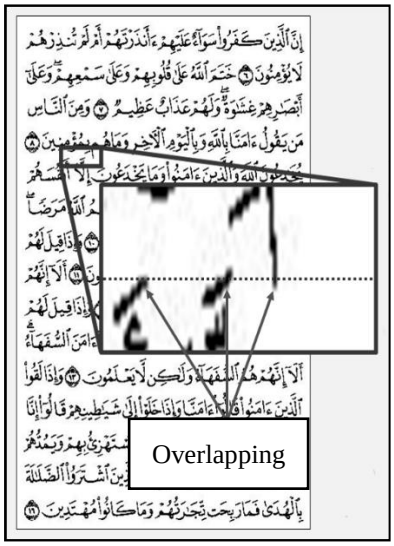


Figure 6: Example image of the overlapping object in parallel horizontal perspective.

Base on the baseline of the text, it specified the number of the text line. The peak for every contour state as its baseline while zero frequency defines as line space. It easy to define space line if the zero frequency exits between two baseline neighbours, it will be defined to be segmented as a row of the text line. But, if there are no zero frequency exits between two baselines neighbours it will determine low contour as space line but need to refined overlapping object. By using this method, it will consider lower contour contains the overlapping object. Thus, overlapping object must be determined by its position by calculating its distance. There is two methods to determine its possession which is based on the distance of baseline and distance of the determined object.

i. Distance of baseline

This find and calculated distance of diacritical marks or stroke of the Arabic word with upper baseline and bottom baseline. The closest baseline is the possession row number of the text line. The overview is illustrated as shown in Figure 7.

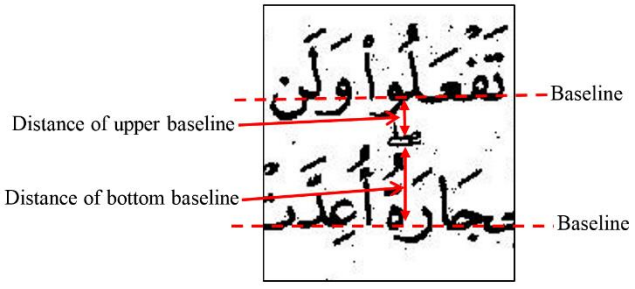


Figure 7: Illustration of the distance of diacritical marks or stroke of the Arabic word with upper baseline and bottom baseline.

ii. Distance of determined object

This find and calculated distance of diacritical marks or stroke of the Arabic word with the closest object that already determined it possession row number of the text line. It will determine its possession row number of line base on its closest object possession.

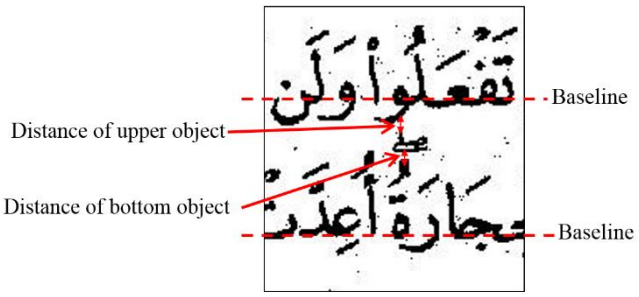


Figure 8: Illustration of the distance of diacritical marks or stroke of the Arabic word with upper baseline and bottom object.

D. Segment Line

The segmentation line will be determined base on horizontal projection profile to detect its number of baseline. Then, it will consider the lower peak of contour as overlap. For overlap, it will determine object possession to determine its row number of the text line. The pseudocode is defined as Figure 9.

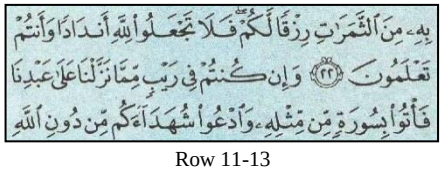
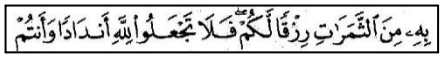
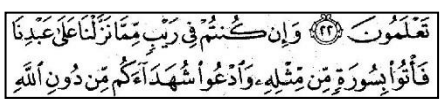
1.0 Start
2.0 Read input image
3.0 Input image → pre-processing image
4.0 Detect baseline using horizontal projection profile
5.0 Fabricate object using neighbouring pixel properties
6.0 Determine object possession
  5.1 Define object possession using a distance of baseline
  5.2 Define object possession using determined object
7.0 Output result image
8.0 End

Figure 9: Illustration pseudocode of proposed method.

IV. Experimental Result

The method was implemented in Java and tested on selected Al-Quran pages. Data set that is being used in this experiment is text image of Mushaf Al-Quran Rasm Uthmani publish by company S Abdul Majeed page 6 row 11-13 and text image of Mushaf Al-Quran from Mushaf Al-Madinah Quran Majeed page 3 row 3-5, row 6-8 and row 8-10. The results are shown in Table 1-4.

Table 1
Result of text image of Mushaf Al-Quran Rasm Uthmani publishes by company S Abdul Majeed page 6.

Input	 Row 11-13
Result of Binary Representation (Laith, 2015) [2]	 Row 11  Row 12-13

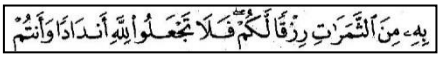
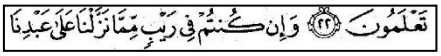
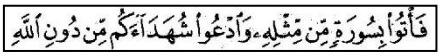
Result of Proposed Method	 Row 11  Row 12  Row 13

Table 2
Result of text image of Mushaf Al-Quran from Mushaf Al-Madinah Quran Majeed page 3.

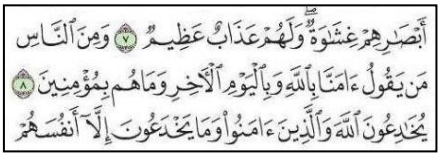
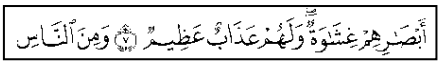
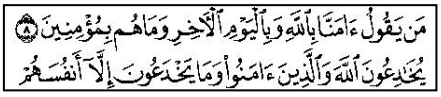
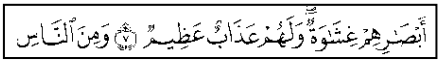
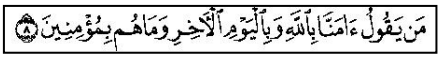
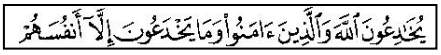
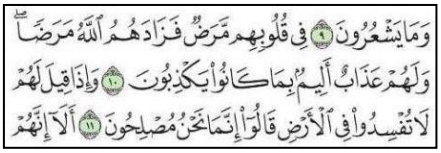
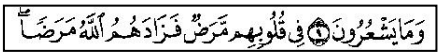
Input	 Row 3-5
Result of Binary Representation (Laith, 2015) [2]	 Row 3  Row 3-5
Result of Proposed Method	 Row 3  Row 4  Row 5

Table 3
Result of text image of Mushaf Al-Quran from Mushaf Al-Madinah Quran Majeed page 3.

Input	 Row 6-8
Result of Binary Representation (Laith, 2015) [2]	 Row 6  Row 7-8
Result of	

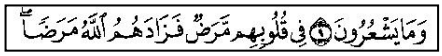
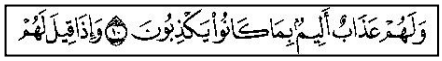
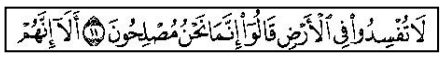
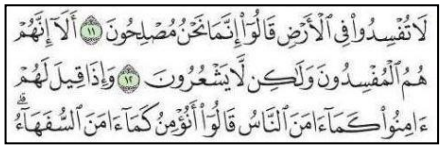
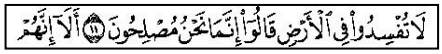
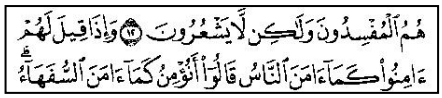
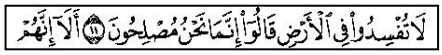
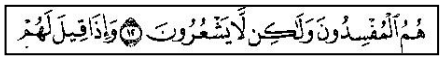
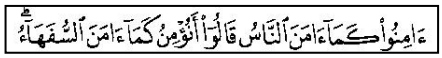
Proposed Method	 Row 6  Row 7  Row 8

Table 4
Result of text image of Mushaf Al-Quran from Mushaf Al-Madinah Quran Majeed page 3.

Input	 Row 8-10
Result of Binary Representation (Laith, 2015) [2]	 Row 8  Row 9-10
Result of Proposed Method	 Row 8  Row 9  Row 10

V. CONCLUSION

In this paper, we proposed an approach for segmenting text line for Mushaf Al-Quran. In addition, this paper describes for identifying overlaps between neighbouring text lines. The purpose of identifying overlaps is by segmenting each line with precision. This paper including the procedure for pre-processing data to form images uniformly to be used throughout the process. However this method not cover for a connected component in the overlap between neighbouring text lines. This is because mostly printed Mushaf Al-Quran does not contain the connected component. If connected component exits, it will allocate base on weightage of distance either upper line or bottom line.

VI. ACKNOWLEDGEMENT

The authors would like to express their appreciation to the Universiti Teknikal Malaysia Melaka for the scholarship of Zamalah UTeM Scheme and funding grant of PJP/2016/FTMK/HI4/S01477 and also the Faculty of Information Technology and Communication for providing the excellent research faculties and facilities.

REFERENCES

[1] C. L. Tan, "Text Line Segmentation for Handwritten Documents Using Constrained Seam Carving," *2014 14th Int. Conf. Front. Handwrit. Recognit. Text*, 2014.

[2] L. N. B. Melhem, "Illumination Removal And Text Segmentation For Al-Quran Using Binary Representation," Thesis for Master, Universiti Teknikal Malaysia Melaka, 2015.

[3] A. Antonacopoulos and D. Karatzas, "Document Image Analysis for World War II Personal Records †," no. January, pp. 336–341, 2004.

[4] J. He and A. C. Downton, "User-Assisted Archive Document Image Analysis for Digital Library Construction," no. Icdar, pp. 1–5, 2003.

[5] U. Pal and S. Datta, "Segmentation of Bangla Unconstrained Handwritten Text," no. Icdar, pp. 3–7, 2003.

[6] A. Zahour, B. Taconet, A. Zahour, B. Taconet, A. Zahour, B. Taconet, S. Ramdane, and P. R. Schuman, "Contribution ` a la segmentation de textes manuscrits anciens To cite this version : Segmentation en ligne Décomposition de l â€™ image en blocs," 2004.

[7] V. Shapiro, G. Gluhchev, and V. Sgurev, "Handwritten document image segmentation and analysis," vol. 14, no. January, pp. 71–78, 1993.

[8] K. L. Banumathi, "Line and word Segmentation of Kannada Handwritten Text documents using Projection Profile Technique," pp. 196–201, 2016.

[9] F. Yin and C. Liu, "Handwritten Chinese text line segmentation by clustering with distance metric learning," *Pattern Recognit.*, vol. 42, no. 12, pp. 3146–3157, 2009.

[10] M. S. Azmi, K. Omar, M. F. Nasrudin, A. K. Muda, and A. Abdullah, "Digital paleography: Using the digital representation of Jawi manuscripts to support paleographic analysis," *Proc. 2011 Int. Conf. Pattern Anal. Intell. Robot. ICPAIR 2011*, vol. 1, no. June, pp. 71–77, 2011.

[11] M. S. Azmi, M. F. Nasrudin, K. Omar, C. W. S. B. C. W. Ahmad, and K. W. M. Ghazali, "Exploiting features from triangle geometry for digit recognition," *2013 Int. Conf. Control. Decis. Inf. Technol. CoDIT 2013*, pp. 876–880, 2013.

[12] M. S. Azmi and K. Omar, "Features Extraction of Arabic Calligraphy using extended Triangle Model for Digital Jawi Paleography Analysis," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 5, pp. 696–703, 2013.

[13] M. S. Azmi, K. Omar, M. F. Nasrudin, A. K. Muda, and A. Abdullah, "Arabic calligraphy classification using triangle model for Digital Jawi Paleography analysis," *Proc. 2011 11th Int. Conf. Hybrid Intell. Syst. HIS 2011*, pp. 704–708, 2011.

[14] A. R. Radzid, "Removing Al-Quran Illumination," Thesis for Bachelor Degree, Universiti Teknikal Malaysia Melaka, 2016.

[15] L. Entrance and M. Perrone, "Pre-processing Methods for Handwritten Arabic Documents Faisal Farooq University at Buffalo," 2005.

[16] H. J. Vala and P. A. Baxi, "A Review on Otsu Image Segmentation Algorithm," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 2, no. 2, pp. 387–389, 2013.

[17] A. Greensted, "Otsu Thresholding." [Online]. Available: http://www.labbookpages.co.uk/software/imgProc/otsuThreshold.html.

[18] B. S. Morse, "Lecture 4 : Thresholding," *Brigham Young University*, 2000. [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MORSE/threshold.pdf.